Applying Fuzzy Clustering algorithm in Genetic Data Set

Ozer OZDEMIR and Asli KAYA

***Abstract*—** Clustering is an important analysis in discovering groups of genes that have similar expression patterns in microarray data analysis. However, the hard clustering method, which correctly assigns each gene to a cluster, is not well appropriate for these data sets. In this work, it is going to be worked with fuzzy clustering algorithm to overcome hard clustering limitations. Fuzzy clustering algorithms allow each gene to belong to all clusters with a certain membership rating by removing the constraint of assigning to a single cluster. It is aimed to perform simpler analyzes by studying fuzzy clustering algorithms, a method of soft techniques developed to meet the need for developing alternative statistical methods in cases where classical clustering analysis is insufficient to analyze data.

***Keywords*—** Clustering, Fuzzy clustering, Fuzzy c-means, genetic.

1. INTRODUCTION

Gene expression data is an invaluable information. However, the natural way to describe patterns that organize the underlying mechanisms of action is to combine genes with similar expression patterns. The first step towards this aim is to adopt a mathematical description of the similarity. Clustering techniques use these mathematical descriptions to group genes in a given sample according to their expression profiles. Clustering algorithms allow each gene to locate the group containing its similar profiles. It is expected that genes in the same cluster have similar biological function. However, biological gene activities are very complex structures. It is known that given genes are subject to regulation by many manners of molecule. The general form of expression of a given gene may therefore correspond to the coincidence of different patterns. To determine this complexity and examine tightly related gene groups, fuzzy clustering algorithms are used that are faster in computing than the classical techniques and contain more flexible capabilities. In contrast to classical (hard) clustering algorithms, fuzzy clustering algorithms allow each gene to be bound to all clusters via a real valued u_{ij} vector. This vector takes values between 0 and 1. The use of membership values u_{ij} helps to identify genes associated with other genes or linked to more than one cluster, thus its biological complexity is discovered.

Ozer OZDEMIR is with the Department of Statistics, Anadolu University Eskisehir 26470 TURKEY (e-mail: ozerozdemir@anadolu.edu.tr).

Aslı KAYA is with the Department of Statistics, Anadolu University Eskisehir 26470 TURKEY (corresponding author to provide phone: 902223350580-4668; e-mail: asli.kaya532@gmail.com).

2. FUZZY CLUSTERING ALGORITHM

Since fuzzy clustering algorithms deal with the uncertainty of real numbers, it helps to reveal clustering patterns that are appropriate for daily life experience. Fuzzy clustering algorithms also use mathematical descriptions, i.e. distance measures, to group similar expressions, such as clustering algorithms. However, unlike classical clustering techniques, each member uses membership functions that allow certain aggregates to be entered into a certain degree. Membership is defined by

$$\begin{aligned} &u_{ij} :\forall i,\forall j \text{ for } i=1,2,...,n,\ j=1,2,...,c \\ &u_{ij} \succ 0 \\ &\sum_{j=1}^{c} u_{ij}=1 \end{aligned} \tag{1}$$

If any data is closer to the cluster center, the membership value of that cluster becomes the largest. The sum of the membership grades of the given word is equal to 1. Fuzzy clustering algorithms usually use the objective function. Objective function based algorithms aim to solve clustering problem by turning it into optimization problem.

2.1. The fuzzy c- means clustering algorithm (FCM)

The most widely used algorithm, based on the least reduction of the objective function, was developed by Bezdek in 1973 [3]. This method is based on the fuzzy logic (1965) proposed by Zadeh [7]. The objective function used in the algorithm is as follows:

$$J(u,\upsilon)=\sum_{j=1}^{n}\sum_{i=1}^{c} u_{ij}{}^{m}\left\|x_j-\upsilon_i\right\|^2 \tag{2}$$

Equation (2) contains a number of variables: there are

$$X=\{x_1,x_2,...,x_n\}\subset R^n = the data, \tag{2a}$$

c = number of clusters in X; $2\le c\le n$, (2b)

m = weighting exponent $1\le m\le\infty$, (2c)

U = fuzzy c- partition of X, (2d)

$\upsilon=\left(\upsilon_1,\upsilon_2,...,\upsilon_c\right)$ = vectors of centers (2e)

$\left\|x_j - \upsilon_i\right\|^2$ is the square of distance from data x_j to centroid υ_i.

$J(u,\upsilon)$ value is the total of pattern measurement of all weighted least square errors. If the objective function is differentiated according to u_{ij} and υ_i, using the Lagrange multipliers method, the obtained equations are as follows;

$$\upsilon_i = \frac{\sum_{j=1}^{n} (u_{ij})^m x_j}{\sum_{j=1}^{n} (u_{ij})^m}, 1 \le i \le c, \quad (3)$$

$$u_{ij} = \left[\sum_{k=1}^{c} \left(\frac{\left\|x_j - \upsilon_i\right\|^2}{\left\|x_j - \upsilon_k\right\|^2} \right)^{1/(m-1)} \right]^{-1}, 1 \le i \le c, 1 \le j \le n. \quad (4)$$

These equations form the iterative optimization process. The FCM algorithm is executed in the following steps:

(A1) Fix c, m, ε and choose any product norm metric for calculation of $\left\|x_j - \upsilon_i\right\|^2$

(A2) Computate of centroids using Eq.(3)

(A3) Update the membership u_{ij} using Eq.(4)

(A4) Repeat (A1) and (A2) until stabilization, i.e. $\left\|U^{(t)} - U^{(t-1)}\right\| \le \varepsilon, t \succ 1$

where ε is the stop criterion between 0 and 1, and t is the number of repetitions. Through this process J converges to a local minimum. The BCO algorithm depends on the randomly initialized values at startup and updates it iteratively using these values. Better performance can be achieved by using an algorithm to identify all centers or by repeatedly running the BCO with different start centers [1].

3. DATA SET

Microarray gene expressions of Alzheimer's disease were obtained from the National Center for Biotechnology Information (NCBI) database, which is open to access. The transformed values of the data were calculated and 2 samples (1 control, 1 severe patient) were run on 22283 genes.

4. ANALYSIS RESULTS

Due to the large size of the data, long-term analysis is not being able to be drawn the graph of the cluster, MATLAB program was studied with. For the fuzzy clustering analysis of the genetic data, the optimum number of clusters was initially obtained. It was analyzed the Fuzzy C- Means algorithm, with all initial parameter values kept constant, the fuzzifier parameter m = 2, and the stop criterion epsilon 1e-6. The most commonly used validity indices were calculated to determine the optimal number of clusters.

Considering the partition coefficient (PC), classification entropy (CE), partition index (SC), separation index (S), Xie and Beni's Index (XB), the optimum number of clusters had been decided. However, evaluation of these two coefficients (PC) and (CE) were abandoned, since it is irrelevant that the number of the clusters for which the minimum value of the classification entropy is usually two or three, the partition coefficient is the maximum and the disconnection of the direct coupling of the partition coefficient to the properties of the data is a disadvantage. In Table 1, the validation indices are given according to the number of clusters.

Küme sayısı	SC	S	XieBeni
2	1.898367203226455, ,	8.519351986835053e-05	1.134104647762376e+02
3	0.943972736536685,	7.658580299960912e-05,	1.365432042175275e+02
4	0.864728925934285, ,	7.327728624217091e-05	1.371314802648906e+02
5	0.762022744075112, ,	5.587579469347124e-05	1.267666874552801e+02
6	0.592283901625667, ,	5.146091812420727e-05	1.234886312021615e+02
7	0.537079379731811,	4.657305480410925e-05,	2.057097602733320e+02
8	0.520511262779214,	4.363853954532626e-05	68.146537980381820
9	0.515699938187614,	4.430969264097327e-05	**49.570149327161590**
10	**0.505575244342024**,	**4.182777497179728e-05**	53.397410887697156

Table 1: Values of Validity Indices for FCM Algorithm

The optimal number of clusters for FCM was selected as 10.

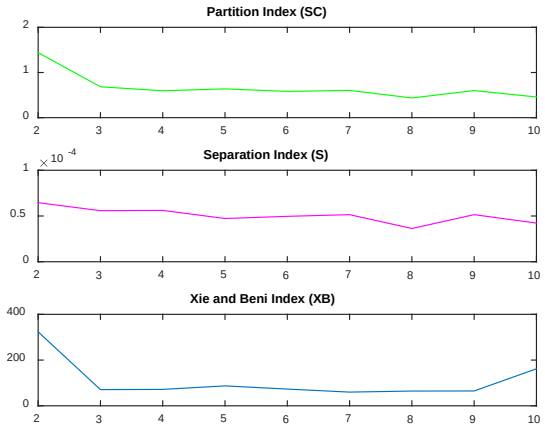

Figure 1: Graphs of Validity Indices for the FCM Algorithm

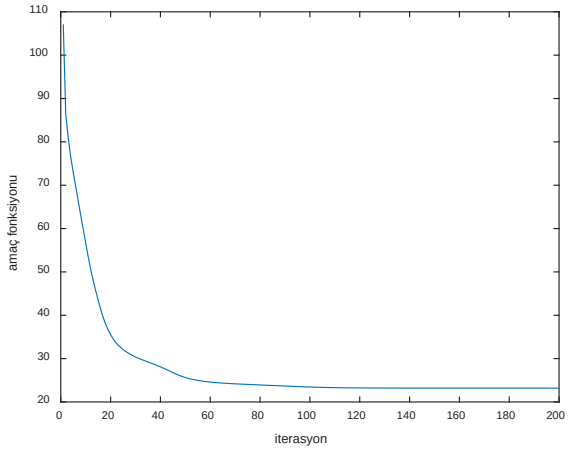

Figure 2: Objective Function Graph

As can be seen from Fig. 2, no significant change in the objective function can be obtained after the 100th iteration.

c	1	2	3	4	5	6	7	8
1	0,2006	0.0716	**0.5058**	0.0071	0.0027	0.0281	0.0538	0.1303
2	8,2900e-05	0.0074	2.5836e-04	1.19225e-05	5.4788-06	3.0149e-05	**0.9911**	0.0011
3	1,1131e-04	0.0136	3.5634e-04	1.5664e-05	7.1624e-06	3,9919e-05	**0.9843**	0.0016
4	0,0039	0,1913	0.0197	4.1112e-04	1.78e-04	0.0012	0.0555	**0.7277**
5	4,7114e-05	0.0046	1.4795e-04	6.7345-06	3.0931e-06	1.7071e-05	**0.09945**	6.4326e-04

Table 2: Membership Values for the FCM Algorithm

As the data set is too large, the membership grades of 5 elements are shown in Table 2 as an example, instead of showing the membership degree of each element. Elements belong to the group they have high membership value. Filled cells show that which element belongs to which cluster.

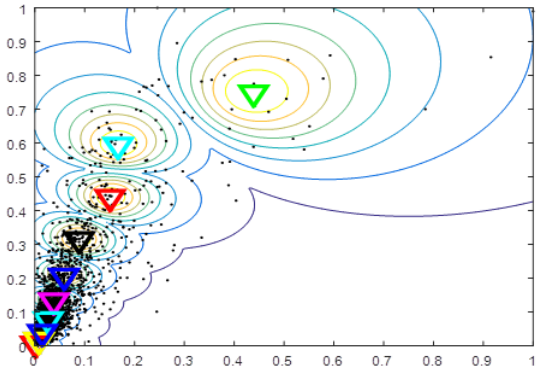

Figure 3: Clustered Data Set with FCM Algorithm

5. Conclusion

Fuzzy clustering is an appropriate method for selecting genes that exhibit a tight association with given clusters. Conventional fractional clustering methods force clusters that do not match all genes to clusters or even variations in expression. Fuzzy set algorithms have been developed as an alternative to hard clustering techniques in high-dimensional data sets such as microarray gene expression data sets. Fuzzy C-means was used in this study aiming to separate groups according to similar expression patterns of gene data from Alzheimer's patients and control group. For this algorithm, all initial parameters are taken as the same, and the optimal clustering numbers required for the algorithm is calculated with the validation indices in the literature. The BCO algorithm has discovered 10 similar patterns. It can be stated that the Fuzzy C- Means method provides a sensitive result. The clusters formed by the FCM algorithm are well separated as shown in Fig.3. The FCM algorithm may try to produce better results for this data. However, it should not be forgotten that the selected initial values are very important for this result.

References

[1] Avcı, U., 2006, Bulanık Kümeleme Algoritmalarının Karşılaştırmalı Analizi ve Bilgisayar Uygulamaları, Yüksek Lisans Tezi, Ege Üniversitesi Fen Bilimleri Enstitüsü, 78 s.

[2] Babuska, R., Fuzzy Systems, Modeling and Identification, Delft University of Technology Department of Electrical Engineering.

[3] Bezdek,J.C. (1981) *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum Press, New York.

[4] Bezdek J. C., "Cluster Validity with fuzzy sets", J. Cybernetics, Vol.3,1974, pp. 58-73

[5] Bezdek, J.C.; Ehrlich R.; Full, W., 1984, FCM: Fuzzy C-Means Algorithm Computers and Geoscience 10 (2-3), 191-203.

[6] Gustafson, D.E., Kessel, W.C., 1979, Fuzzy Clustering with a Fuzzy Covariance Matrix, IEEE CDC San Diego, 761-766.

[7] Höppner, F., Klawonn, F., Rudolf, K., Runkler, T., 1999, Fuzzy Cluster Analysis: Methods for Classification Data Analysis and Image Recognition, John Wiley & Sons, p. 5-75.

[8] Wang, W., Zhang, Y., 2007, On Fuzzy Cluster Validity İndices, Fuzzy Sets and Systems, 158, 2095-2117.

[9] Zadeh, L.A., 1965, Fuzzy Sets, Information and Control, 8, 338-353.

[10] http://www.mathworks.com/access/helpdesk/help/toolbox/fuzzy/

[11] https://www.ncbi.nlm.nih.gov/geo/

Ozer Ozdemir was born in Turkey in 1982. He received his B.Sc., M.Sc. and Ph.D. degrees in statistics in the Department of Statistics at Anadolu University, Turkey, respectively in 2005, in 2008 and in 2013. He has worked as a Research Assistant from 2006-2008, as a Lecturer from 2008-2014 and as an Assistant Professor from 2014 in the Department of Statistics at Anadolu University, Turkey.
He has published over 50 international conference papers and journals in his research areas. His research interests include Applied Statistics, Simulation, Artificial Neural Networks, Fuzzy Logic, Fuzzy Modeling, Time Series, Computer Programming, Statistical Software and Computer Applications in Statistics.

Aslı Kaya was born in Turkey in 1991. She received her B.Sc. degree in Statistics in the Department of Statistics at Anadolu University, Turkey, in

2014. She received her B.Sc. degree in the Department of Business in Faculty of Management at Anadolu University, Turkey, in 2015. She still continue her education as a master student in the Department of Statistics at Anadolu University, Turkey from 2015.